Daily et al. BMC Bioinformatics 2011, 12:495
http://www.biomedcentral.com/1471-2105/12/495

BMC Bioinformatics

RESEARCH ARTICLE Open Access

MotifMap: integrative genome-wide maps of regulatory motif sites for model species

Kenneth Daily[1,2], Vishal R Patel[1,2], Paul Rigor[1,2], Xiaohui Xie[1,2] and Pierre Baldi[1,2,3*]

Abstract

Background: A central challenge of biology is to map and understand gene regulation on a genome-wide scale. For any given genome, only a small fraction of the regulatory elements embedded in the DNA sequence have been characterized, and there is great interest in developing computational methods to systematically map all these elements and understand their relationships. Such computational efforts, however, are significantly hindered by the overwhelming size of non-coding regions and the statistical variability and complex spatial organizations of regulatory elements and interactions. Genome-wide catalogs of regulatory elements for all model species simply do not yet exist.

Results: The MotifMap system uses databases of transcription factor binding motifs, refined genome alignments, and a comparative genomic statistical approach to provide comprehensive maps of candidate regulatory elements encoded in the genomes of model species. The system is used to derive new genome-wide maps for yeast, fly, worm, mouse, and human. The human map contains 519,108 sites for 570 matrices with a False Discovery Rate of 0.1 or less. The new maps are assessed in several ways, for instance using high-throughput experimental ChIP-seq data and AUC statistics, providing strong evidence for their accuracy and coverage. The maps can be usefully integrated with many other kinds of omic data and are available at http://motifmap.igb.uci.edu/.

Conclusions: MotifMap and its integration with other data provide a foundation for analyzing gene regulation on a genome-wide scale, and for automatically generating regulatory pathways and hypotheses. The power of this approach is demonstrated and discussed using the P53 apoptotic pathway and the Gli hedgehog pathways as examples.

Background

A central challenge of biology is to map and understand gene regulation on a genome-wide scale. For any given genome, only a small fraction of the regulatory elements embedded in the DNA sequence have been characterized, and there is great interest in developing computational methods to systematically map all these elements and understand their relationships. Such computational efforts, however, are significantly hindered by the overwhelming size of non-coding regions and the statistical variability and complex spatial organizations of regulatory elements and interactions, especially in mammalian species.

While many gene-specific, condition-specific, and factor-specific resources for motif binding sites exist [1-4], it is perhaps surprising that genome-wide systematic catalogs of binding sites for most species do not. Past efforts have focused primarily on the yeast and fly genomes and with severe restrictions, for instance in terms of data (e.g. ChIP-seq only) or genomic regions (e.g. promoter only). The prototype MotifMap system [5] used an improved comparative genomics approach to provide one of the first genome-wide maps for the human genome and test its accuracy. This system, however, has several limitations including the direct use of coarse genome alignments for searching for binding sites leading to missed and incorrectly scored sites, and the unavailability of maps for other model species. Furthermore, while the available lists of transcription factors are not exhaustive, new information about transcription factors and regulatory interactions is continuously being produced and thus such maps must be periodically updated.

* Correspondence: pfbaldi@ics.uci.edu
[1]Department of Computer Science, University of California Irvine, Irvine, CA 92697 USA
Full list of author information is available at the end of the article

Here we describe improvements to the prototype methods that are used with a new whole-genome alignment and an expanded list of transcription factors to create a new, more comprehensive, map for the human genome. Furthermore, we apply the updated methodology to the genomes of other model organisms for which alignments and estimated phylogenetic trees are available, creating genome-wide maps for the yeast, worm, fly and mouse genomes.

At its core, MotifMap uses data from transcription factor binding motif databases, specifically JASPAR [6] and TRANSFAC [7]. For yeast and fly, we have supplemented the matrices available from JASPAR and TRANSFAC with those available from a number of publications (see Additional file 1 for a full list of the sources for each species). The binding matrices are used to search a reference genome for binding sites and produce three scores at each site. The first score is the Normalized Log-Odds (NLOD) score derived from the position weight matrix of the corresponding transcription factor. The second score is the Bayesian Branch Length Score (BBLS) to measure the degree of evolutionary conservation. Functional elements, such as those playing a regulatory role, often evolve more slowly than neutral sequences and can be detected by their higher level of conservation. MotifMap uses publicly available whole genome alignments and the corresponding phylogenetic trees to leverage the power of comparative genomics in order to eliminate false positive hits. The third score is the False Discovery Rate (FDR) estimated by using Monte Carlo methods. The three scores at each site are used, in combination with other filters, to generate genome-wide maps.

The quality of the maps is assessed and compared against our previous results [5] as well as other methods [8,9] in various ways, including comparison to experimental data, such as high-throughput ChIP-seq data. The maps provide a foundation for inferring regulatory networks and can be integrated with a variety of other heterogeneous and autonomous data sources.

Methods

Normalized Log-Odds score (NLOD)

Binding sites for each transcription factor are identified by scanning the genome sequence with a position weight matrix. We transform each original weight matrix into a log-odds matrix to account for the background frequency of the nucleotides across the genome. The log-odds score of a sequence is computed as

$$\mathrm{LOD}(S) = \sum_{j=1}^{|S|} f(x)$$

Where

$$f(x) = \begin{cases} \log_2(x) & \text{if } q_{ij} > eb_i 2^c \\ \frac{x}{e2^c \log(2)} + c & \text{if } q_{ij} \le eb_i 2^c \end{cases}$$

where $x = \frac{q_{ij}}{b_i}$, the value q_{ij} from the position weight matrix is the probability of observing nucleotide i({A, C, G, T}) at position j in a sequence S of length $|S|$, and b_i is the probability of observing nucleotide i in the entire genome. For reasonable values of q_{ij} corresponding to $x > e2^c$, the function is simply equal to $log_2(x)$. However, for small values of q_{ij} corresponding to $x \le e2^c$, the logarithm function can take large negative values. Traditionally, to avoid this problem, pseudocounts are added to the frequency matrices, in a heuristic and matrix-dependent fashion. The alternative approach proposed here lower bounds the values of each scoring matrix directly by replacing the log function around zero with a continuous linear approximation. In this work, we use c = -3.

The motif matching score is scaled to fall between 0 and 1 to yield the normalized log-odds score:

$$\mathrm{NLOD}(x) = \frac{\mathrm{LOD}(x) - y_{\min}}{y_{\max} - y_{\min}}$$

where y_{max} and y_{min} are the maximum and minimum LOD scores that the matrix can achieve by using the most likely or least likely nucleotide at each position. A z-score is also derived from the NLOD score by estimating the mean and variance of the score of random sequences across the genome. For mammalian species, we use a z-score threshold of 4.27, corresponding to a p-value of 0.00001, to find a list of initial candidate sites across the reference genome. For yeast, fly, and worm, we use a lower threshold corresponding to a z-score between 2.57 and 3.72, or a p-value between 0.005 and 0.0001. Finally, we restrict the total number of binding sites by ordering the sites for each motif individually by their z-score, and keeping sites with a z-score at least as high as the k^{th} site. For our purposes, k = 100,000, as was done in the prototype version.

Bayesian Branch Length Score (BBLS)

Many previous methods have shown that evolutionary conservation can be used to identify transcription factor binding sites [10-12]. An innovative aspect of the MotifMap system is how the degree of evolutionary conservation is assessed using the Bayesian Branch Length Score (BBLS) [5], which itself is an improvement over a previous score, the Branch Length Score (BLS) [13,14]. More precisely, given a multiple alignment of N species and their evolutionary tree, a transcription factor motif, and the genome coordinates of a candidate binding site,

let σ_i = 0 or 1 denote the presence or absence of the motif at the aligned location in the corresponding species i. The BLS is simply the total length of the branches associated with the most recent common ancestor of all the species for which σ_i is set to 1. However, in reality σ_i is not a binary variable but rather comes with a probability p_i measuring the degree of confidence in whether the corresponding motif is present or not in species i at the corresponding location. Given a set of N aligned species, the BBLS takes into account this uncertainty by computing the *expected BLS* in the form:

$$\begin{aligned} BBLS &= E(BLS) \\ &= \sum_{\sigma_1,\ldots,\sigma_N} P(\sigma_1,\ldots,\sigma_N) BLS(\sigma_1,\ldots\sigma_N) \\ &= \sum_{\sigma_1,\ldots,\sigma_N} r_1\ldots r_N BLS(\sigma_1,\ldots\sigma_N) \end{aligned} \quad (1)$$

Where

$$r_i = \begin{cases} p_i \text{ if } \sigma_i = 1 \\ 1 - p_i \text{ if } \sigma_i = 0 \end{cases}$$

The values of p_i for the leaves of the tree are derived using the NLOD score described above. If the corresponding z-score is too low, p_i is set to 0. An efficient dynamic programming approach, avoiding the addition of an exponential number of terms (Equation 1), has been derived [5], and a corresponding software implementation is available (see below).

False Discovery Rate (FDR)

For every motif weight matrix, we generate control matrices by randomly shuffling the columns of the motif weight matrix. The shuffling is repeated up to 10,000 times so as to produce up to 10 control matrices. The shuffled matrices must be sufficiently different from the original one to be used as control matrices. In practice, we use a cutoff of 0.35 on the similarity measure computed by first taking the average correlation between columns over pairs of windows of length 8 in the original and permuted motif, then taking the maximum of these correlations over all pairs of windows, and then normalizing by the length of the motif. Only binding matrices are retained that: (1) are at least eight nucleotides long; and (2) can produce at least three sufficiently different shuffled versions for the Monte Carlo FDR procedure. In addition, for mammalian species, each shuffled matrix is restricted to have the same CG-dinucleotide frequency as the original matrix. The same motif searching procedure is used with each control matrix. The False Discovery Rate is computed as the median number of sites found using the shuffled matrices divided by the number of sites found for the real matrix at a particular (NLOD, BBLS) score combination or higher.

Sequence alignments and modular design

The prototype version of MotifMap searched the low-resolution multiple alignment files obtained from the UCSC Genome Browser [15] directly. As a result, possible alignments of a motif could be missed in other species, for example in poorly aligned regions with many gaps. To address this problem, the overall methodology used to search for aligned transcription factor binding sites has been considerably improved (Figure 1).

The new approach searches instead the reference genome directly and uses the low-resolution alignments only as a seed to identify regions in other species aligning to the motif in the reference species. An expanded sequence including 15 base pairs on each side of each binding site in the reference species is used to identify aligned regions in the other genomes. This expanded sequence helps compensate for the low-resolution nature of the whole genome alignments [16]. Furthermore, instead of using the aligned regions directly, which may be too short or contain many gaps, we find all the alignment blocks overlapping the expanded sequence. Due to the nature of the algorithm used to build the multiple alignments, the sequences in different aligning blocks for any single species may be very far apart from each other on the chromosome, or even on completely separate chromosomes. As a result, we only concatenate blocks that are within 30 base pairs of each other (maintaining any intervening sequence). This operation yields a set of blocks of aligning regions; each block contains sequences from other species aligned to the binding site. For each species, we find the motif sequence with the highest normalized log-odds score across all blocks. Finally, the scores corresponding to the selected sequence from each species are used for BBLS scoring. In practice, requiring a minimum number of species to be aligned to the reference sequence at each binding site improves performance. The default requirement, used for instance in the case of the yeast map, is set to at least one other species (i.e. BBLS >0). For the human map, in the public version of MotifMap, binding sites are required to be conserved across at least four non-primate species. This also enables a fair comparison to the prototype version that used the same requirement.

Because the new modular design of MotifMap is not dependent on searching the UCSC coarse multiple alignment files directly, it enables one to also use other alignments if necessary, such as high resolution alignments of the upstream regions of known homologous or orthologous genes, even when these are not in the UCSC format (e.g. the MAF format produced by the multiz alignment software), or to focus the search on

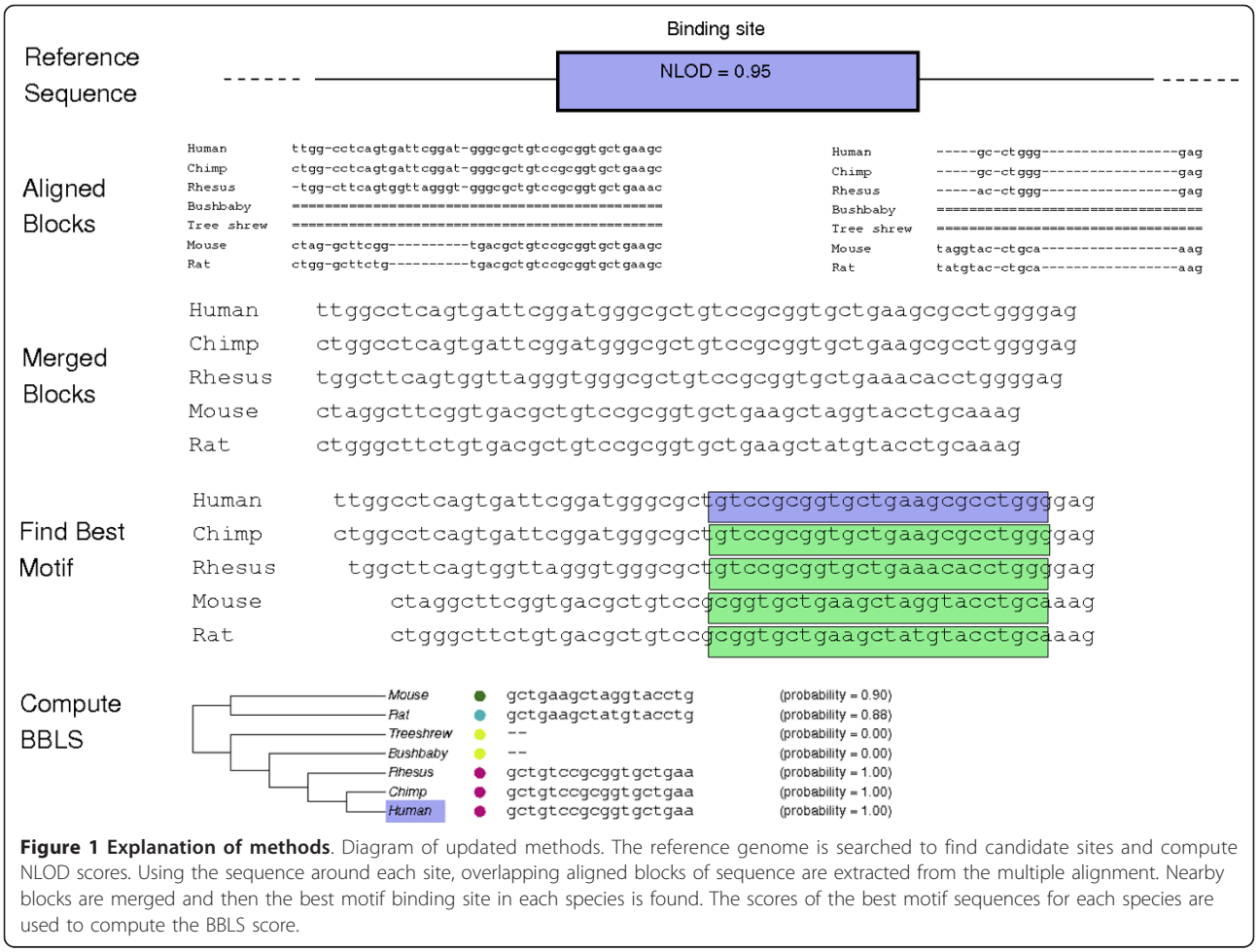

Figure 1 Explanation of methods. Diagram of updated methods. The reference genome is searched to find candidate sites and compute NLOD scores. Using the sequence around each site, overlapping aligned blocks of sequence are extracted from the multiple alignment. Nearby blocks are merged and then the best motif binding site in each species is found. The scores of the best motif sequences for each species are used to compute the BBLS score.

any subset of the genome. To avoid bias from binding sites that occur in regions that are conserved for being part of a translated portion of a gene and are not necessarily under positive selection because of their importance for regulatory control, we exclude exonic regions of the genome from the default public version of MotifMap. Likewise, we exclude repetitive regions.

Redundancy filter

A transcription factor is often annotated with multiple binding matrices in JASPAR and TRANSFAC. For example, each matrix may represent a specific isoform of the factor dependent on the biological context (e.g. cell type or experimental condition). However, in order to estimate a total number of unique potential binding sites, a given site can be counted only once for a given transcription factor, even when this factor has multiple binding matrices. For this purpose, we first perform the genome-wide search independently for each matrix, and then group overlapping binding sites. We choose a representative for each transcription factor in that group by picking the site with the highest BBLS score. The final result is a non-overlapping, non-redundant, list of binding sites for each transcription factor.

Results

New MotifMaps

Each MotifMap is generated automatically via a pipeline running on a parallel computer cluster. Comprehensive maps for human, mouse, fly, worm, and yeast have been generated and new maps can be produced automatically. Details about the genomes, alignments, and matrices used in each MotifMap can be seen in Table 1. The raw data for the total number of binding sites across the genomes ranges from hundreds of thousands for yeast, worm, and fly to millions for mouse and human. Table 2 summarizes the number of transcription factors, matrices, and binding sites for each available species after all filtering steps have been applied. For the human MotifMap, we predict 519,108 binding sites for 570 matrices, nearly a 5-fold increase over the number of sites and matrices in the prototype version, while maintaining a low FDR of 0.1 or less.

Table 1 Multiple alignment information

Species	Build	Alignment	# species in alignment	# of matrices
Yeast	sacCer2	multiz7way	7	507
Worm	ce6	multiz6way	6	6
Fly	dm3	multiz15way	12 (files only)†	262
Mouse	mm9	multiz30way	30	830
Human	hg18	multiz28way	17 (placentals only)†	837
Human	hg19	multiz46way	32 (placentals only)†	837

Number of species in multiple alignments for each reference species, and the number of original matrices for each species. Comprehensive maps exist for yeast, worm, fly, mouse, and human (both for a 17-species and 32-species alignment). See the UCSC Genome Browser website for more information on the alignments used and the species they contain. †Trees used are truncated to remove the most distant species.

Evaluation of new methods using experimental data

We first compare the updated methodology to the prototype version using data on well-studied transcription factors and experimentally-determined binding sites using high-throughput methods, such as ChIP-seq. While ChIP-seq and related methods are not perfect, they still provide the best available experimental approximations to genome-wide maps of binding sites. While the prototype map used 17 species, a larger number of genomes and genome alignments has become available since its publication. Thus, for comparison purposes, we run the new methodology using both the same tree of 17 species used for the first prototype, as well as an expanded tree containing 32 placental mammals.

Specifically, we consider the same set of highly studied transcription factors (Table 3), same motifs, same experimental data [17-22], and same whole genome alignments as in Xie et al. [5], to compute the area under the Receiver Operating Characteristic (ROC) curves (AUC) using the updated methodology. For all motifs, we see an improvement of the AUC in the range of 1-5% over the previous version. [Note that when computing the AUC, we include all ChIP-seq regions that do not contain a conserved motif binding site in the class of true negatives, as in [5]. However, we still robustly observe improvements in the range of 0-5% when not including these regions in the class of true negatives.] For P53, CTCF, and NRSE, we observe an increase in the AUC with a decrease in the number of sites found. For NFKB and STAT1, we observe a modest increase in the number of sites along with an increase in the AUC. We also observe further modest improvements for a few of these transcription factors when the number of species in the multiple alignments is increased from 17 to 32 placental mammals (see the UCSC Genome Browser website for details on the species in each alignment).

Table 2 Non-redundant transcription factor binding sites

Species	# Transcription Factors	# Matrices	# Sites	# Sites FDR ≤ 0.1
Yeast	161	147	115,387	1,577
Worm	6	6	88,895	69
Fly	94	66	191,655	36,091
Mouse	473	575	6,617,325	740,685
Human (hg18)	468	570	2,554,732	519,108
Human (hg19)	468	530	1,410,309	457,198

Number of non-redundant transcription factors and binding sites across the genome (see text for definition of "non-redundant").

Table 3 Performance comparison of the prototype and updated MotifMap pipelines

	NFKB	MYC	P53	STAT1	CTCF	NRSE
	AUC					
Prototype, 17 species	0.722	0.683	0.861	0.606	0.814	0.941
Update, 17 species	**0.797**	0.765	**0.902**	0.780	0.887	0.950
Update, 32 species	0.786	**0.812**	0.896	**0.820**	**0.903**	**0.951**
	Number of sites					
Prototype, 17 species	11,636	55,271	28,635	6,134	69,446	13,055
Update, 17 species	13,924	100,311	24,880	9,537	53,794	7,488
Update, 32 species	14,839	100,275	25,563	10,034	77,064	8,127

Area under the ROC curve (AUC) and number of sites found for the prototype MotifMap pipeline versus the updated MotifMap pipeline (performed with the original 17-species alignment and a newer 32-species alignment); the best performing method for each motif is shown in bold.

We also use ChIP-seq data available for 35 mouse transcription factors obtained from the TRANSFAC suite to further assess the performance of the MotifMap pipeline and compare it with other methods. We evaluate the performance of the BBLS scoring scheme to recover known binding sites identified by ChIP-seq against four other scores: BLS [13,14], NLOD (as described in this work), PhastCons [8], and PhyloP [9]. Each score is individually used to rank the binding sites identified by MotifMap. We calculate the number of true and false positive sites identified in the ChIP-seq data to compute the AUC, as in Xie et. al. [5]. Table 4 summarizes the results for the performance of the MotifMap pipeline in recovering the sites identified by the ChIP-seq methods by reporting the results for the 20 top transcription factors with the largest AUC values. For these 20 transcription factors, we see performances comparable to those seen for the human MotifMap: MotifMap achieves the best AUC result in 16 of them, while relatively small differences (3% or less) are seen for the remaining four, providing further evidence of the overall quality of the MotifMap system and its ability to generalize and identify binding sites in other species.

Localization analysis: binding site location properties

To further assess the quality of the maps, we examine the distribution of the candidate sites relative to the

Table 4 Performance of the mouse MotifMap

Name	BBLS	BLS	NLOD	PhastCons	PhyloP
Ctcf	**0.901**	0.838	0.893	0.798	0.754
Myc:Max	**0.831**	0.826	0.731	0.773	0.690
Zfp281	**0.827**	0.820	0.611	0.691	0.679
Tcfcp211	**0.809**	0.500	0.754	0.800	0.668
c-Myc	**0.778**	0.771	0.734	0.758	0.710
Gli3	0.772	0.771	0.619	**0.806**	0.659
Gli1	**0.770**	0.728	0.727	0.704	0.689
E2f5	**0.760**	0.737	0.632	0.737	0.667
Myc	**0.760**	0.699	0.540	0.703	0.718
Pdx1	0.757	**0.765**	0.500	0.696	0.689
Trim28	**0.753**	0.749	0.609	0.640	0.642
Klf4	**0.740**	0.500	0.500	0.695	0.678
Esrrb	**0.739**	0.500	0.516	0.667	0.608
Zfa	**0.733**	0.731	0.660	0.677	0.644
Mycn	**0.730**	0.728	0.620	0.690	0.664
Cnot3	0.683	**0.688**	0.568	0.614	0.597
Stat3	**0.677**	0.634	0.656	0.655	0.614
Ppara	**0.673**	0.664	0.642	0.636	0.615
Nr0b1	**0.668**	0.653	0.598	0.612	0.597
Zfp42	0.629	0.627	0.596	0.650	**0.661**

Area under the ROC curve (AUC) for predicting transcription factor binding sites identified by ChIP-seq experiments in mouse. Each column is associated with a different method for scoring and ranking the putative sites identified by MotifMap, from which ROC curves and AUCs are computed. The best performing method for each motif is shown in bold.

locations of genes across the genome. Using the high confidence data (FDR ≤ 0.1), we find that the majority of sites are within 1 Kbp of the transcription start sites (TSS) of known genes across all species. Figure 2 shows a plot of the distribution of distance to the closest gene for each binding site for the human genome. This distribution becomes increasingly peaked as one increases the BBLS threshold filter (Figures 3a, b). However, we note that we also find high-confidence sites significantly far from known transcription start sites (further than 100 Kbp away). These sites would be missed in a promoter-only analysis of transcription factor binding sites. We see similar distributions for mouse, while for smaller genomes (such as yeast and fly) the binding sites are even closer to the transcription start sites. This is expected, since the genomes of these species are more condensed, including shorter promoter and intragenic regions.

MotifMap system, web server, and data integration

The MotifMap "system" consists of three main components: (1) a computational pipeline to perform the genome-wide search; (2) a database to store candidate motif binding sites, the scores associated with them, and the relationships to other features; (3) custom code to interface between the database and a web service; and (4) a Flex web application, to display data to users. All steps in the pipeline for identifying and scoring binding sites are performed in parallel using a high performance computer cluster. Along with the locations and scores for each binding site, we compile and store relationships between the binding sites and other genomic features, such as genes (RefSeq [23] and Ensembl [24]) and Gene Ontology (GO) annotations [25]. Some species (fly and yeast) use specific gene annotation resources instead (FlyBase [26] and SGD [27]). The database is currently being expanded as other MotifMaps and new relationships become available. The binding site data and relationships for all available species are publicly available through the MotifMap web site (http://motifmap.igb.uci.edu).

While the prototype MotifMap version had a simple interface to display data, the new web application has been extensively upgraded with multiple features and functionalities to allow users to explore these genome-wide datasets more easily. User can interactively select a model species and one or more transcription factors, visualize the logos of the corresponding motifs, filter the results by various criteria and thresholds (genome location, NLOD/z-score, BBLS, FDR), and retrieve a corresponding list of binding sites, with the distances to the nearest TSS and the corresponding gene annotations. The results can be downloaded in a variety of standard formats (GFF, BED, CSV) or exported directly for visualization in the UCSC Genome Browser. Furthermore, for each motif binding site, users can view the local multiple alignment and the phylogenetic tree with the corresponding probability scores for each species, as shown in simplified form at the bottom of Figure 1. A Python implementation of an efficient algorithm for computing the Bayesian Branch Length Score can also be downloaded from the MotifMap web site. MotifMap uses an integrative approach combining, for instance, phylogenetic, genomic, and transcription factor data. The resulting maps themselves can in turn be integrated with many other datasets (see Discussion). Two kinds of data that are fully integrated into the MotifMap database and available to the user are GO annotations and SNPs. For instance, for a given GO annotation and the corresponding set of genes, user can retrieve all the nearby candidate binding sites. Likewise, SNPs falling within or near a transcription factor binding site have the potential for influencing the regulation of the corresponding gene [28]. Thus it is useful to be able to list which SNPs in a GWAS (Genome Wide Association Study) or other genotyping study fall within or nearby transcription factor binding sites. Analyses of GWAS data focused primarily on coding regions run the risk of missing important SNPs affecting regulatory regions. The relationship between SNPs and binding sites has

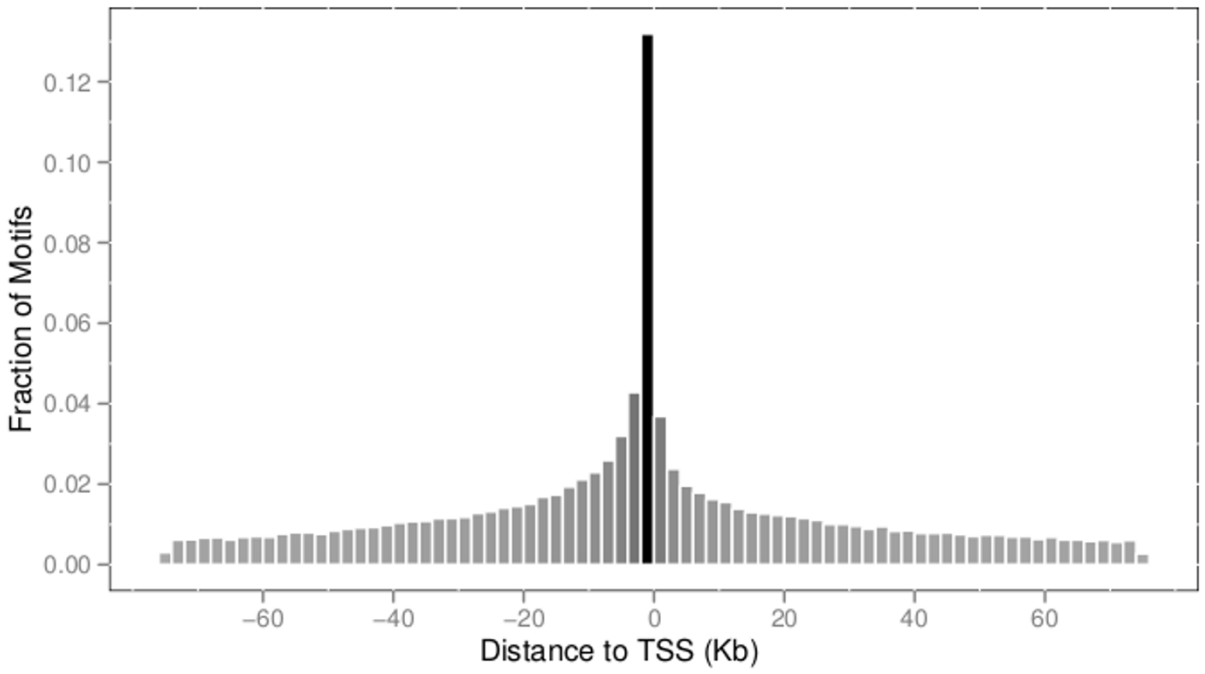

Figure 2 Distribution of distance to closest gene for human binding sites. Distribution of the distance to the closest gene (Transcription Start Site or TSS) for high confidence human motif binding sites.

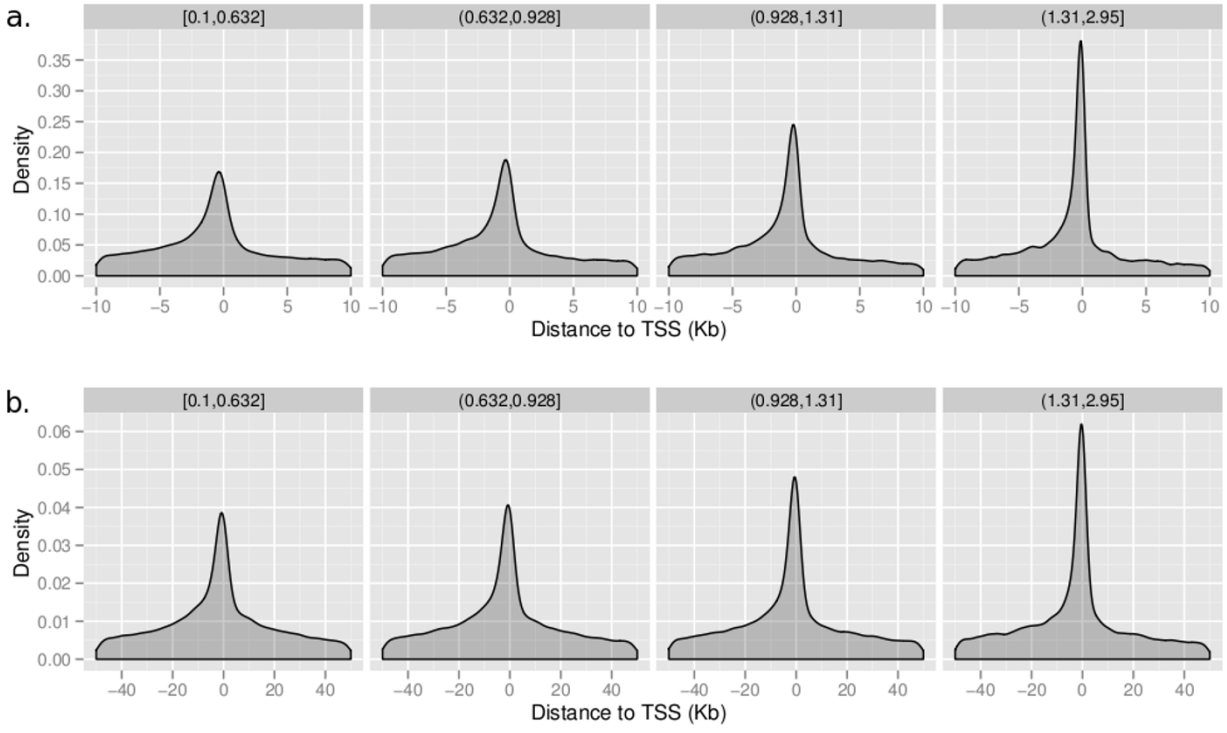

Figure 3 Distribution of MotifMap regulatory elements as a function of conservation. Empirical distribution of distances of human transcription factor binding sites to the closest (≤ 10 Kbp and ≤ 50 Kbp) RefSeq gene transcription start site (TSS). The sites are grouped into quartiles according to the BBLS score; each group has one quarter of the total binding sites. The BBLS range for each quartile is given at the top of each plot. As the BBLS conservation score increases, we observe a larger proportion of binding sites close to the TSS of the closest gene.

been integrated into the MotifMap web application as an additional analysis tool called SNPer, which allows the retrieval of motif binding sites that overlap with SNP sites. The HapMap3 [29] and dbSNP [30] datasets are currently available for use with the mouse and human MotifMap. Users can download the MotifMap results for further integration with specific GWAS or other studies.

Discussion

The MotifMap approach has allowed us to derive state-of-the-art genome-wide maps of candidate regulatory elements for some of the main model organisms, in particular for mouse and human. For the worm, the map produced is considerably more primitive because only six transcription factor binding matrices are available in TRANSFAC and JASPAR. However, the availability of the map for this limited set of transcription factors may still be of some use and all the maps will be updated as more binding matrices become available.

Each binding site predicted by MotifMap corresponds in fact to a regulatory hypothesis, thus a single MotifMap can generate from thousands to millions of hypotheses. These hypotheses can be tested and refined in the laboratory, either individually in the case of very specific interactions which can be tested with great precision, or on a larger but less precise scale using high-throughput methods, such as ChIP-seq. These multiple hypotheses can also be further refined and analyzed by computational methods using integrative approaches where regulatory hypotheses are simultaneously combined: (1) with each other in the form of regulatory networks; and (2) with other kinds of data. Regulatory hypotheses can be integrated with each other to identify regulatory networks of transcription factors, including regulatory loops and, for instance, hypothesize that transcription factor A regulates transcription factor B, transcription factor B regulates transcription factor C, and transcription factor C regulates transcription factor A. These networks and loops can be thought of as the core regulatory network of a cell. Regulatory hypotheses can also be integrated with many other kinds of data to refine regulatory inferences, as described in the Results section using GO and SNP data and below with other kinds of data. In particular, MotifMap and GO annotations can be used to infer the common functions of a set of genes targeted by a transcription factor or, conversely, to infer the transcription factor that may regulate a set of genes with common GO annotations. To illustrate these ideas, here we give a simple demonstration of the power of integrating MotifMap and other data to generate regulatory network hypotheses, above the level of an individual regulatory site. For demonstration purposes, we choose two examples. We reconstruct the P53 apoptotic pathway, since it is an important and well-studied pathway which allows us to assess the quality of the predictions. We also apply the same general ideas to the Gli family of transcription factors and the hedgehog pathway to demonstrate the effectiveness of these methods on a relatively less-studied transcription factor family and pathway where important regulatory effects remain to be discovered.

Mouse P53 apoptotic pathway

We attempt to reconstruct the P53 direct regulatory interactions in the mouse P53 apoptotic pathway using data from MotifMap for putative P53 binding sites across the genome. We first compile a list of over 380 unique gene transcripts from the RefSeq database [23] annotated with the Gene Ontology term "Apoptosis" (GO:0006915). We then retrieve predicted P53 binding sites from MotifMap in the promoter region of these genes to generate a regulatory network of P53's role in apoptosis. The promoter region of a gene is defined as 15 Kbp upstream and 3 Kbp downstream, which approximately encompasses the region associated with the first intron, from the transcription start site. To evaluate the network generated from MotifMap data, we compare it to the P53 pathway described in the KEGG database [31], which reports 14 genes directly regulated by P53 in the apoptotic pathway (Figure 4). Table 5 shows the number of known and potentially novel P53 targets predicted as a function of FDR. At a FDR of 0.05, we predict eight target genes from the list of all apoptotic genes, six of which are annotated in KEGG. Searching the literature reveals that the other two target proteins, DDIT4 and PHLDA3, are also known targets of P53 [32,33] but not annotated in KEGG. At a FDR of 0.25, we predict a total of 71 targets, including 12 of the 14 targets annotated in KEGG; the only exceptions are FAS and TSAP6 (also called STEAP3). FAS is a predicted direct target, but has a slightly higher FDR (0.28). For TSAP6 we find two P53 sites (1784 bp and 4582 bp upstream) with a strong motif matching score; however these sites are not conserved. A novel predicted target is BID, which is annotated in KEGG as a downstream indirect target in the P53 apoptotic pathway. If we reduce the length of the upstream promoter regions from 15 Kbp down to 5 Kbp, the same KEGG targets are recovered with the exception of PIDD and SHSA5. A few targets have P53 binding sites downstream of the TSS, in the first intron, and these would not have been recovered with a search focused on promoter regions only. Thus in short the MotifMap system is capable of robustly recovering most of the direct targets of P53 described in KEGG, as well as providing a ranked list of potential new targets, some of which can be confirmed by a literature search.

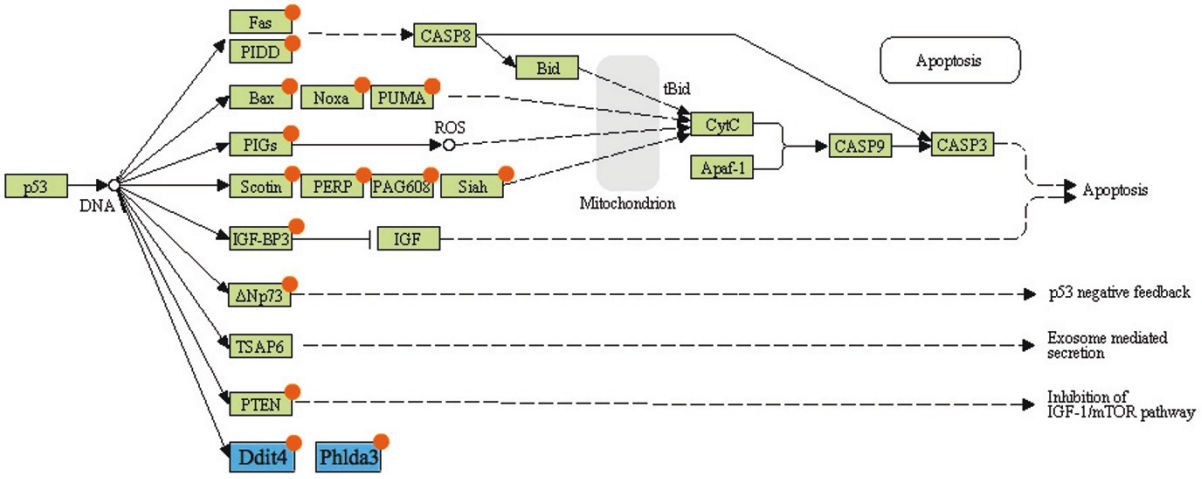

Figure 4 Known apoptotic targets of P53. Known apoptotic genes from the KEGG pathway database and the literature for P53. Genes in light green are annotated in KEGG. Orange dots indicate direct targets recovered by MotifMap. DDIT4 and PHLDA3 are examples of additional direct targets identified by MotifMap with FDR < 0.05 which have been reported in the literature but are not present in KEGG.

Mouse Gli hedgehog pathway

Next, we examine the Gli family of transcription factors. Although Gli is a relatively less studied transcription factor, mutations in Gli genes have been associated with multiple developmental disorders and cancers [34]. We first compile a list of Gli targets. The KEGG database lists only two annotated targets of Gli1 (Hhip and Ptch1), as well as an autoregulatory loop of Gli1. Gli1 is annotated as a downstream effector of the Sonic hedgehog pathway [34]. In addition, Gli1 is known to regulate the Wnt signaling pathways [35]. Due to the lack of many annotated targets in KEGG, we used the Transcriptional Regulatory Element Database (TRED) [36], which contains an additional four annotated Gli family targets. We find Gli binding sites predicted by MotifMap in the promoter region of the seven annotated targets and also many of the Wnt proteins. We observe predicted binding sites in the Shh promoter (14,843 bp upstream) as well as in the second intron. In addition, we recover the Gli1 autoregulatory loop [37] and regulation of Gli3 by Gli1 [38] (Figure 5a). All binding sites for all targets are recovered at an estimated FDR ≤ 0.25, within 15 Kbp upstream and 3 Kbp downstream of each gene. Furthermore, we identify a highly conserved binding site (BBLS >7, perfectly conserved in 27 out of the 30 species in the alignment) near Ptch1. Nkx2-8 and Nkx2-2, both of which have been reported as targets of Gli family transcription factors [39,40], have predicted binding sites within 2 Kbp upstream of the transcription start site with similar conservation (Figure 5b). We also identify Rab34 as a true Gli target [39] at a lower conservation level (BBLS >2); this threshold includes approximately 100 novel targets.

Table 5 Mouse P53 apoptotic pathway

FDR	0	0.05	0.1	0.15	0.2	0.25
KEGG known	4	6	6	7	10	12
Potentially novel	1	2	16	29	50	73
Total	5	8	22	36	60	85

Number of known (annotated in KEGG) and potentially novel P53 direct targets predicted at different FDR thresholds.

Further integration and challenges

Regulatory networks do not consist only of transcription factors and their direct regulatory interactions, but can include also protein-protein interactions (PPI). Integrating PPI (physical or genetic) data [41,42] with protein-DNA interactions from MotifMap can yield a more comprehensive view of molecular mechanisms and networks. Integration of PPI data can also facilitate the identification of transcriptional complexes. For example, evidence for a complex based on adjacency of binding sites for two transcription factors could be strengthened by data supporting physical interactions between these factors. In general, however, factors with proximal binding sites need not physically interact with each other in order to influence transcription, and MotifMap data can be used to identify modules of transcription factors with co-occurring binding sites near co-regulated genes. To

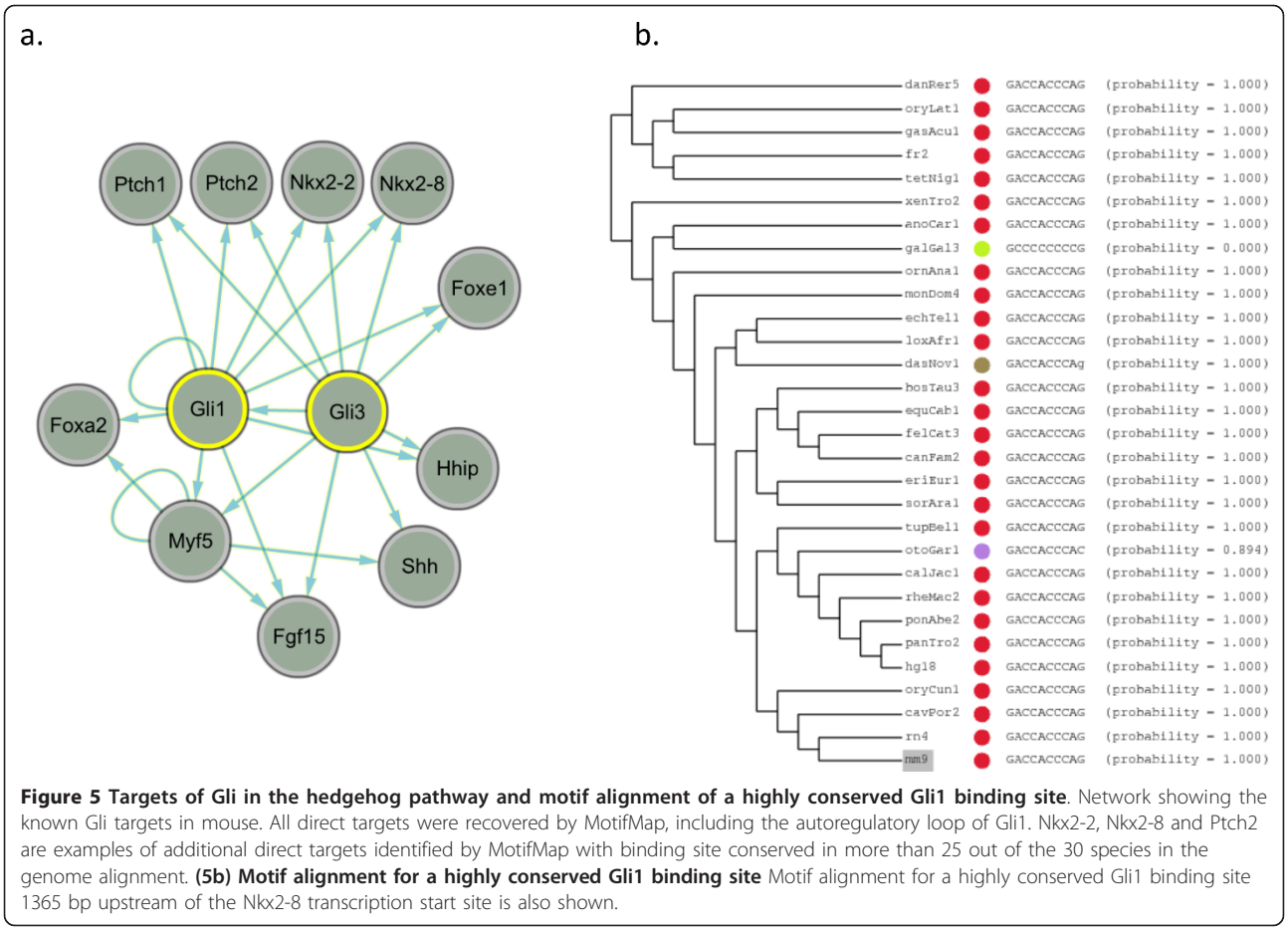

Figure 5 Targets of Gli in the hedgehog pathway and motif alignment of a highly conserved Gli1 binding site. Network showing the known Gli targets in mouse. All direct targets were recovered by MotifMap, including the autoregulatory loop of Gli1. Nkx2-2, Nkx2-8 and Ptch2 are examples of additional direct targets identified by MotifMap with binding site conserved in more than 25 out of the 30 species in the genome alignment. **(5b) Motif alignment for a highly conserved Gli1 binding site** Motif alignment for a highly conserved Gli1 binding site 1365 bp upstream of the Nkx2-8 transcription start site is also shown.

derive a more accurate and complete global picture, it is also important to incorporate information about RNA elements involved in gene regulation [43]. As so far described, MotifMap provides a static view of potential transcription-factor/DNA interactions. Since transcription factor regulation of most genes does not occur ubiquitously or constantly across all cells in an organism, DNA microarrays and high-throughput sequencing of transcripts (RNA-seq) provide another important source of information about the cell-specific, tissue-specific, or condition-specific expression of genes. Thus MotifMap can be integrated with gene expression data, such as the Gene Expression Omnibus (GEO) data [44]. This integration provides additional information about, for instance, the average direction of a particular interaction (up- or down-regulation) across many experiments, or about the specific portion of the total potential regulatory network that is activated in a given condition. An important challenge ahead lies in better understanding the role of epigenetics in the regulation of gene transcription. An interesting source of data for further integration with MotifMap comes from the ENCODE project [45] providing the locations of epigenetic signatures, such as histone tail methylations or acetylations, across the human genome for a large number of cell lines. Combinations of these markers can identify transcription factor binding sites that are specific to a particular cell line; for example, the presence of H3K4Me1 and absence of H3K4Me3 denotes enhancer regions. This integration induces regulatory sub-networks, potentially describing important interactions needed for a particular cell type to function properly.

Another considerable challenge is the role of chromatin and 3D structure in gene regulation. New high-throughput techniques like Chromosome Conformation Capture-on-Chip (4C), Hi-C and Chromatin Interaction Analysis using Paired-End Tag sequencing (ChIA-PET) allow the detection of long range or inter-chromosomal interactions of DNA [46-48]. This provides the ability to detect regulatory elements that may be distal to the gene they regulate linearly, but are brought close together in 3-dimensional space. For instance, a recent study used 4C to investigate the properties and dynamics of the genomic loci that are in contact with glucocorticoid receptor (GR) responsive loci [49]. Incorporating this kind of data into MotifMap could provide further evidence of these distant regulatory

interactions and improve our ability to infer regulatory mechanisms and networks.

Many other data, such as scientific literature, or information about diseases and drugs, are also being integrated in house with MotifMap. Each data comes with its own noise and limitations and it is the combination of diverse lines of evidence that has the power to solidify inferences and rank hypotheses in a relevant way. This integration process is not new, of course, and in essence is at the root of IBM's Watson system for the game of Jeopardy [50]. This integration process is ongoing and raises computational challenges both in its execution and in what can be served publicly given a limited amount of computational resources.

Finally, another potential computational challenge for systems like MotifMap is the dynamic use of evolutionary trees and comparative genomics. The current version of MotifMap builds a genome-wide map, assessing conservation with a single static tree for each species. But clearly not all regulatory elements are conserved, and even when they are, the optimal tree for assessing their degree of conservation may vary with each transcription factor and each biological question. Thus studying how to dynamically assess conservation, including its weaker forms [51,52], and how to discover regulatory elements that are poorly conserved remain important questions for further investigations.

Conclusion

The MotifMap system aims to provide comprehensive genome-wide map of regulatory elements for each organism. Since experimental data on gene expression obtained with DNA microarray or high-throughput sequencing methods is inherently biased (to a specific condition, cell type, etc.), a resource that catalogs transcription factor binding sites across the entire genome in an unbiased fashion is valuable. We have created the first such comprehensive maps of candidate regulatory motifs across the yeast, fly, worm, mouse, and human genomes. The updated methodology has improved the detection of experimentally validated motif binding sites and, together with integration with other data, the generation of regulatory networks and hypotheses. Overlaying and integrating information from multiple sources, well beyond transcription factor binding motifs and genomic DNA sequences, is key to building better maps and ultimately to understanding gene regulation on a genome-wide scale.

Additional material

Additional file 1: Sources of binding matrices. Table listing the original source of each transcription factor binding matrix.

Acknowledgements

This work was in part supported by National Institutes of Health grants LM010235-01A1 and 5T15LM007743 and National Science Foundation grant MRI EIA-0321390 to PB, and by the UCI Institute for Genomics and Bioinformatics. We also wish to thank NVIDIA for hardware support.

Author details

[1]Department of Computer Science, University of California Irvine, Irvine, CA 92697 USA. [2]Institute for Genomics and Bioinformatics, University of California Irvine, Irvine, CA 92697 USA. [3]Department of Developmental and Cell Biology, University of California Irvine, Irvine, CA 92697 USA.

Authors' contributions

PB conceived the study and the algorithms and coordinated and supervised all aspects. XX contributed to the algorithms and the coordination. KD, VP, and PB wrote the manuscript. PR, VP, and KD wrote the software and implemented the web server. KD, VP, and PB performed the detailed analyses. All authors proofread and approved the final manuscript.

Received: 29 September 2011 Accepted: 30 December 2011
Published: 30 December 2011

References

1. Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E: **AGRIS: the Arabidopsis Gene Regulatory Information Server, an update.** *Nucleic Acids Research* 2011, **39(suppl 1)**:D1118-D1122.
2. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS: **REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila.** *Nucleic Acids Research* 2010, **39(suppl 1)**:D118-D123.
3. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJM, Consortium TORA: **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Research* 2008, **36(suppl 1)**:D107-D113.
4. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG: **Transcription Regulatory Regions Database (TRRD): its status in 2002.** *Nucleic acids research* 2002, **30**:312-317[http://www.ncbi.nlm.nih.gov/pmc/articles/PMC99088/].
5. Xie X, Rigor P, Baldi P: **MotifMap: a human genome-wide map of candidate regulatory motif sites.** *Bioinformatics* 2009, **25(2)**:167-174.
6. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic acids research* 2010, , **38 Database:** D105-110[http://dx.doi.org/10.1093/nar/gkp950].
7. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DUU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic acids research* 2003, **31**:374-378[http://dx.doi.org/10.1093/nar/gkg108].
8. Siepel A, Bejerano G, Pedersen J, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier L, Richards S, *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome research* 2005, **15(8)**:1034-1050.
9. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Research* 2010, **20**:110-121[http://dx.doi.org/10.1101/gr.097857.109].
10. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: **The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates.** *Genome Biology* 2005, **6(12)**:R104.
11. Elemento O, Tavazoie S: **Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach.** *Genome biology* 2005, **6(2)**:R18.
12. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters**

and 3' UTRs by comparison of several mammals. *Nature* 2005, **434(7031)**:338-345.

13. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby GG, Brennecke J, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SWW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent JJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.** *Nature* 2007, **450(7167)**:219-232.
14. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.** *Proceedings of the National Academy of Sciences* 2007, **104(17)**:7145-7150.
15. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010.** *Nucleic acids research* 2010, , **38 Database:** D613-619[http://dx.doi.org/10.1093/nar/gkp939].
16. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner.** *Genome Research* 2004, **14(4)**:708-715.
17. Johnson D, Mortazavi A, Myers R, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316(5830)**:1497.
18. Wei C, Wu Q, Vega V, Chiu K, Ng P, Zhang T, Shahab A, Yong H, Fu Y, Weng Z: **A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome.** *Cell* 2006, **124**:207-219[http://dx.doi.org/10.1016/j.cell.2005.10.043].
19. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, *et al*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nature methods* 2007, **4(8)**:651-658.
20. Zeller KI, Zhao X, Lee CWH, Chiu KP, Yao F, Yustein JT, Ooi HS, Orlov YL, Shahab A, Yong HC, Fu Y, Weng Z, Kuznetsov VA, Sung WK, Ruan Y, Dang CV, Wei CL: **Global mapping of c-Myc binding sites and target gene networks in human B cells.** *Proceedings of the National Academy of Sciences* 2006, **103(47)**:17834-17839.
21. Lim C, Yao F, Wong J, George J, Xu H, Chiu K, Sung W, Lipovich L, Vega V, Chen J, *et al*: **Genome-wide mapping of RELA (p65) binding identifies E2F1 as a transcriptional activator recruited by NF-κB upon TLR4 activation.** *Molecular cell* 2007, **27(4)**:622-635.
22. Kim T, Abdullaev Z, Smith A, Ching K, Loukinov D, Green R, Zhang M, Lobanenkov V, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128(6)**:1231-1245.
23. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Research* 2009, **37(suppl 1)**:D32-D36.
24. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJP, Parker A, Proctor G, Vogel J, Searle SMJ: **Ensembl 2011.** *Nucleic Acids Research* 2011, **39(suppl 1)**:D800-D806.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**:25-29[http://dx.doi.org/10.1038/75556].
26. Drysdale R, t FC: **FlyBase Drosophila.** In *Methods in molecular biology (Clifton, N.J.), Volume 420 of Methods in Molecular Biology*. Edited by: Dahmann C, Walker JM, Walker JM. Totowa, NJ: Humana Press; 2008:45-59 [http://dx.doi.org/10.1007/978-1-59745-583-1_3].
27. project S: **Saccharomyces Genome Database.** *Saccharomyces Genome Database* 2011 [http://downloads.yeastgenome.org/].
28. D'Souza UM, Craig IW: **Functional polymorphisms in dopamine and serotonin pathway genes.** *Human Mutation* 2006, **27**:1-13[http://dx.doi.org/10.1002/humu.20278].
29. International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426(6968)**:789-796.
30. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucl Acids Res* 2001, **29**:308-311[http://dx.doi.org/10.1093/nar/29.1.308].
31. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28**:27-30[http://dx.doi.org/10.1093/nar/28.1.27].
32. Ellisen LW, Ramsayer KD, Johannessen CM, Yang A, Beppu H, Minda K, Oliner JD, McKeon F, Haber DA: **REDD1, a Developmentally Regulated Transcriptional Target of p63 and p53, Links p63 to Regulation of Reactive Oxygen Species.** *Molecular Cell* 2002, **10(5)**:995-1005[http://www.sciencedirect.com/science/article/pii/S1097276502007062].
33. Kawase T, Ohki R, Shibata T, Tsutsumi S, Kamimura N, Inazawa J, Ohta T, Ichikawa H, Aburatani H, Tashiro F, Taya Y: **PH Domain-Only Protein PHLDA3 Is a p53-Regulated Repressor of Akt.** *Cell* 2009, **136(3)**:535-550 [http://www.sciencedirect.com/science/article/pii/S0092867408015638].
34. Matise MP, Joyner AL: **Gli genes in development and cancer.** *Oncogene* 1999, **18(55)**:7852-7859.
35. Mullor JL, Dahmane N, Sun T, Ruiz i Altaba A: **Wnt signals are targets and mediators of Gli function.** *Current biology: CB* 2001, **11(10)**:769-773[http://view.ncbi.nlm.nih.gov/pubmed/11378387].
36. Jiang C, Xuan Z, Zhao F, Zhang MQ: **TRED: a transcriptional regulatory element database, new entries and other development.** *Nucleic acids research* 2007, , **35 Database:** D137-D140[http://dx.doi.org/10.1093/nar/gkl1041].
37. Weiner HL, Bakst R, Hurlbert MS, Ruggiero J, Ahn E, Lee WS, Stephen D, Zagzag D, Joyner AL, Turnbull DH: **Induction of Medulloblastomas in Mice by Sonic Hedgehog, Independent of Gli1.** *Cancer Research* 2002, **62(22)**:6385-6389[http://cancerres.aacrjournals.org/content/62/22/6385.abstract].
38. Hu MC, Mo R, Bhella S, Wilson CW, Chuang PT, Hui Cc, Rosenblum ND: **GLI3-dependent transcriptional repression of Gli1, Gli2 and kidney patterning genes disrupts renal morphogenesis.** *Development* 2006, **133(3)**:569-578.
39. Vokes SA, Ji H, McCuine S, Tenzen T, Giles S, Zhong S, Longabaugh WJR, Davidson EH, Wong WH, McMahon AP: **Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning.** *Development* 2007, **134(10)**:1977-1989.
40. Santagati F, Abe K, Schmidt V, Schmitt-John T, Suzuki M, Yamamura Ki, Imai K: **Identification of Cis-regulatory Elements in the Mouse Pax9/Nkx2-9 Genomic Region: Implication for Evolutionary Conserved Synteny.** *Genetics* 2003, **165**:235-242[http://www.genetics.org/content/165/1/235.abstract].
41. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human Protein Reference Database-2009 update.** *Nucl Acids Res* 2009, **37(suppl_1)**:D767-772.
42. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucl Acids Res* 2006, **34(suppl_1)**:D535-539.
43. He L, Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation.** *Nature Reviews Genetics* 2004, **5(7)**:522-531.
44. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucl Acids Res* 2009, **37(suppl_1)**:D885-890.
45. Consortium TEP: **A User's Guide to the Encyclopedia of DNA Elements (ENCODE).** *PLoS Biol* 2011, **9(4)**:e1001046+[http://dx.doi.org/10.1371/journal.pbio.1001046].
46. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).** *Nature Genetics* 2006, **38(11)**:1348-1354.

47. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.** *Science* 2009, **326(5950)**:289-293.
48. Fullwood MJ, Wei CL, Liu ET, Ruan Y: **Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.** *Genome Research* 2009, **19(4)**:521-532.
49. Hakim O, Sung MH, Voss TC, Splinter E, John S, Sabo PJ, Thurman RE, Stamatoyannopoulos JA, de Laat W, Hager GL: **Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements.** *Genome Research* 2011, **21(5)**:697-706.
50. Ferrucci D: **Build Watson: an overview of DeepQA for the Jeopardy! challenge.** *Proceedings of the 19th international conference on Parallel architectures and compilation techniques* PACT '10, New York, NY, USA: ACM; 2010, 1-2[http://doi.acm.org/10.1145/1854273.1854275].
51. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.** *Science (New York, NY)* 2010, **328(5981)**:1036-1040.
52. King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, groups for Transcriptional Regulation E, Analysis MS, Chiaromonte F, Miller W, Hardison RC: **Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data.** *Genome Research* 2007, **17(6)**:775-786.

doi:10.1186/1471-2105-12-495
Cite this article as: Daily *et al.*: MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics* 2011 **12**:495.